1 AGENCY USE ONLY *(Leave blank)*	2 REPORT DATE	3 REPORT TYPE AND DATES COVERED
	May 1991	Professional Paper

4 TITLE AND SUBTITLE

SOME PRAGMATIC ISSUES OF MEASUREMENT

5 FUNDING NUMBERS

In-house

6 AUTHOR(S)

T. P. Enderwick

7 PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Ocean Systems Center
San Diego, CA 92152-5000

8 PERFORMING ORGANIZATION REPORT NUMBER

9 SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Naval Ocean Systems Center
San Diego, CA 92152-5000

10 SPONSORING/MONITORING AGENCY REPORT NUMBER

11 SUPPLEMENTARY NOTES

12a DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b DISTRIBUTION CODE

A-1

13 ABSTRACT *(Maximum 200 words)*

Measurement is the cornerstone of Human Factors (HF) research and testing. To facilitate discussions in this paper HF testing will be treated as a special case of HF research in that testing uses many of the same methods and measurements.

HF research is applied research which means the results are always expected to have a use. This does not mean that the results are necessarily unrelated to theory evaluation. This simply means that the sponsor and user have the right to know what the utility and limitations of the results are in relation to the specific problems or questions posed prior to the research. This, in turn, determines the selection and/or development of measures to be used in research and testing. This gives rise to a number of pragmatic issues for research in general and measurement in particular.

The topics in this paper are somewhat diversified. First, some of the potential users will be identified along with their needs in respect to HF research and testing. This will be followed by a discussion of some pragmatic issues and end with a suggested approach for evaluating crew's contribution to system performance.

Published in *Proceedings of Human Factors Society 34th Annual Meeting*, Oct 1990.

14 SUBJECT TERMS			15 NUMBER OF PAGES
			16 PRICE CODE
17 SECURITY CLASSIFICATION OF REPORT	18 SECURITY CLASSIFICATION OF THIS PAGE	19 SECURITY CLASSIFICATION OF ABSTRACT	20 LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAME AS REPORT

21a. NAME OF RESPONSIBLE INDIVIDUAL	21b. TELEPHONE *(Include Area Code)*	21c. OFFICE SYMBOL
T. P. Enderwick	(619) 553-8007	Code 443

SOME PRAGMATIC ISSUES OF MEASUREMENT

Thomas P. Enderwick
Naval Ocean Systems Center
San Diego, California

The pragmatic issues in this paper are varied. The topics cover the users of human factors measures, various ways of how we report human factors findings, and a suggested approach for assessing the operator's contribution to system performance.

INTRODUCTION

Measurement is the cornerstone of Human Factors (HF) research and testing. To facilitate discussions in this paper HF testing will be treated as a special case of HF research in that testing uses many of the same methods and measurements.

HF research is applied research which means the results are always expected to have a use. This does not mean that the results are necessarily unrelated to theory evaluation. This simply means that the sponsor and user have the right to know what the utility and limitations of the results are in relation to the specific problems or questions posed prior to the research. This, in turn, determines the selection and/or development of measures to be used in research and testing. This gives rise to a number of pragmatic issues for research in general and measurement in particular.

The topics in this paper are somewhat diversified. First, some of the potential users will be identified along with their needs in respect to HF research and testing. This will be followed by a discussion of some pragmatic issues and end with a suggested approach for evaluating crew's contribution to system performance.

HUMAN FACTORS USERS

Users of HF Results

The potential users of HF research results do not constitute a homogeneous group in that their individual data and information needs are different from one another. Below are some of the users and my brief interpretation of theirs HF needs.

Researcher. Here we are talking about a researcher who is reviewing others' research. Some of his/her needs include the research findings related to his/her current problem to avoid duplications and reported mistakes. He/she is also looking for improved or novel methods and measures. Ideally, he/she will use the information to design and conduct his/her research/tests to (1) answer an immediate research question and (2) meet the needs of other users who will be briefly described below.

Tester. The person conducting a system or equipment test can have a somewhat more limited scope than that of the researcher in that the tester's job is to determine whether or not established HF standards and/or criteria have been met through system and equipment design, manpower allocations, personnel selection, and training. The measures used in testing are

based upon the selected standards and criteria that are specific to the system or equipment being evaluated.

System designer. Some of his or her needs include learning what the limitations and capabilities are of those individuals who will man the system, what stability and what variability the individuals will contribute to system performance, and which of those capabilities must be recruited and developed through training. The performance measures are needed to satisfy the designer's needs.

Equipment designer The equipment designer's needs include the traditional human factors engineering data of MIL-STD- 1472D as well as current research findings that supplement existing HF engineering data. It is the individual man-machine interface that is his/her concern. Many of the measures required here are physiological and anthropological measures.

Manager. There are some occasions when managers of systems need HF data. For example, the system manager in considering the expandion of the systems capability by adding another sensor, asks "Do I have to add another technician to the crew to maintain the system?" He/she wants data that can support a decision one way or another. The HF job is to determine the relative use of the technicians' work hours. How much of the technicians' time is spent on system maintenance and how much is spent on non-system related activities? Time measures are the obvious HF end product.

Training course developers. They will use research and test data to establish training priorities, and as a bases for evaluating performance during and after training.

Contract specialists. These appear somewhat remote to HF except in a legal sense. However, they prepare contract specifications and negotiate provisions the contracts which can have HF implications. If realistic measures can be specified, the HF provisions are easier to prepare and to defend. To be effective the measures and their implications need to be understood by the specialist. This is a responsibility of the HF specialist.

Expanding the Usability of Measures

There may be other users as well as those discussed above. It is not either possible or practical to take measures of research or testing and analyze them to meet others' needs. However, I would like to propose a way that we can improve the utility of the measures we do take for our own purposes.

It is based upon the fact that research is designed to collect performance data under specified conditions using specific measures. When data collection is completed measurements are transformed and/or re-arranged for the purposes of analysis. This formatted data is then subject to selected analyses. The analyses are selected so the results will provide answers to the specific research questions. Other analyses could be performed on these data but there is often neither time or funds available to do them. It is recommended that the formatted data be made available for other potential users to analyze. The formatted data would be a supplement to the report that describes the conditions, etc. under which they were collected. The supplement publication costs

could be passed on to the requester.

MEASUREMENT AND ANALYSIS ISSUES

Precision: Values and Pitfalls

Laboratory vs. field precision. The measures used in laboratory and simulators are usually more precise than those used in the field. However, this does not mean that laboratory measures are more useful than those used in the field. The reason is that the research or testing performed in the laboratory and in the field have different objectives. It really is a matter of the question(s) being asked.

For example, in the laboratory we may be concerned with the human ability to discern "just noticeable differences"(JNDs) in light levels, i.e., how many light intensity units will produce a JND for the subject? The results may show that N units produced a JND in the mid-range with N+1 units to produce a JND in the upper range and perhaps N-1 units to have a JND in the lower range.

Now take a situation in an airplane cockpit where there are different intensity control knobs for each of several sets of instrument displays. The objective is to determine the number of discrete settings the knobs need to have. The cockpit is outfitted with photometers to measure and record the ambient light. The pilot takes off at sunset and flies until after full darkness sets in.

If JNDs were used as the perceptual units of measure to determine number of intensity settings that would be needed for adequate illumination of the instruments, the resulting data would be misleading. An actual study was conducted for that purpose by Rockwell International for the Air Force in the late 60s. The pilots were allowed to change the intensity setting at any time during the flight. The results showed that they changed the setting only once or twice between sunset and darkness. Although the laboratory measure was more precise, it would have been inappropriate in the operational setting since visual perception was only one factor in the operation of the aircraft. Precision should not be the sole criteria in selecting measures for any purpose.

Levels of precision. There are times when a measure is acceptable but the size of the units (precision) is inappropriate. An paleoarchaeologist may be satisfied with time estimates of a prehistoric event given as occurring "500 million years ago, give or take 100 thousand years". This can be contrasted to measuring computer operations where nanoseconds are too large.

An example of selecting an appropriate measure with inappropriate units was evident in an unpublished draft report. One of a number of measures used in the evaluation of two sets of controls/displays was that of time. Time was an appropriate measure as were the other measures, but the time units were too small.

The report covered the test and evaluation of two identical back hoes; one with a standard controls/displays arrangement and the other with a non-standard arrangement. Digging trenches was one of the tasks used in the test. The testers appropriately measured the depth and width of the trenches to the nearest inch (a desireable level of precision when digging

around gas pipes and foundations), but then measured the speed of operations in seconds. Seconds are not meaningful in heavy equipment operations. At most such operations are measured in fractions of hours and more often in hours. The mean times in second were significantly different for using the different arrangements only if seconds were used in the analysis. Again, having more precision is not always desirable and can produce misleading results.

Using Measures in Analysis

Measures in and measures out of analysis. In research and testing we have a tendency to take performance measures of a number of individuals doing the same tasks under specified conditions and integrating these measures by calculating their means and standard deviation. Both of these calculations are usually found in the subsequent report and sometimes in HF standards. These values may be readily usable by researchers and testers, but not necessarily by deigners and engineers.

Designers and engineers usually need the HF parameter estimates to incorporate into their designs. Would it not be more useful to provide them with a range of values to work with based upon the range covered by the standard deviations. For example, we could provide them with the mean as the point estimate and the range end values as the plus and minus tolerances. The number of deviations from the mean would be used to establish the range. The number of standard deviations would depend upon the particular measure and the possible uses of the measure by the designer.

Statistical vs. practical significance. The difference between statistical and practical significance is known by those who perform statistical analyses and many of those who use the statistical results. I question, however, whether or not all the users of our research results know what statistical significance means and how to distinguish between it and practical significance.

Any elementary text on statistics will operationally define statistical significance as having a large enough difference between means that we would expect such a difference to occur by chance alone, on the average and in the long run, once out of every 100 times we ran the identical experiment (the .01 level). This is recognized as an arbitrary and accepted standard established by the research community for its own use.

In contrast, practical significance is based on how you are going to use the statistical results. Anderson, in his succinct book "The Psychology Experiment" aptly described the difference between statistical and practical significances by first giving a similar definition to statistical significance as that above. For illustrating practical significances he used an example of having a barrel filled with 100 pistols with only five of the pistols having a bullet in the chamber. When you reach into the barrel you have a .05 chance of picking a loaded gun. As Anderson put it, how much confidence would you have that you gun is loaded if you were going to use it in a duel? On the other hand, how much confidence would you have that the gun is loaded if you were to play Russian roulette?

A more realistic example of the difference between statistical and practical significance can be found in medical research where a drug may have proved to be effective in remitting a fatal condition in only 5 out of 10,000 patients. This would not be a statistically significant finding - except for those five patients. It could be possible, in this case, that the effectiveness of the drug could be linked to the patients' individual body chemistry rather than chance. If so, the medical research community ought to be studying the individuals to determine why they responded rather than discard the results as a chance happening.

A CREW/SYSTEM EVALUATION APPROACH

Estimating the potential for improvement

The navy has what is known as Land Base Test Sites (LBTSs) which are high fidelity simulators of shipboard systems. At present they have two uses; first, for engineering design and system testing, and second, for crew training. These LBTSs would be ideal for crew research and testing. All inputs to crew stations are controllable via the use of computers. The system outputs are also recorded with the aid of the computers. Realistic operating conditions can be manipulated to satisfy a wide array of research and test designs. Crew/system responses can be recorded with little, if any, intrusions into the crew operations. The approach to be described requires such a simulator.

It occurred to me during an operational test at one of these LBTSs that the LBTS could be used to derive an estimate of the remaining potential for crew improvement. The method to do this is described below with the idea of stimulating discussion.

The basic assumption is that members of an "ideal" crew are able to perform their assigned tasks without error and within time requirements. There would be no potential for crew improvement. If an ideal crew existed, then their performance measures could be compared to those of a real crew. The difference between the performance of the ideal crew and that of the real crew would be the potential for real crew improvement. Of course there are no such ideal crews, but one can be simulated.

The ideal crew is simulated by having the testing staff guide the real crew through their assigned mission tasks. This minimizes the times for a crew member to perceive the inputs and take appropriate actions.

Prior to the crew being "idealized", they would first carry out the same mission tasks without guidance and the system performance would be recorded. The same mission would then be repeated with the simulated ideal crew. The guidance to the crew would be as detailed and timely as possible. For example, the member of the testing staff would tell the crew member where on the scope the target will appear, when it will appear, and what to do when it does, i.e., no errors and done on time. Again the system performance is recorded.

The comparison to be made would be between the system performance measures generated by the "real" crew and by the "ideal" crew. The difference between the two system performance scores would be the estimate for potential

system performance improvement through crew performance improvement. If the the difference were small, then further efforts to improve system performance through crew improvement would not be cost effective. If the difference were reasonably large it would require further inquiry as to what was contributing to poor crew performance on that mission, e.g., inadequate training, poor control-display layout, or poor operating procedures.

This same scenario could be repeated using different missions. This would provide the means for an overall evaluation of the crews relative contribution to system performance. If this approach can be used successfully in simulators with similar characteristics, we will have one more tool to measure the crews' effect on system operations.